MASTER'S THESIS

Impact van een Geautomatiseerde Feedbacktool op de Studie-ervaring en Leeropbrengst.

Schokkelé, William

Award date:
2020

Impact van een Geautomatiseerde Feedbacktool op de Studie-ervaring en Leeropbrengst.

Impact of an Automated Feedback Tool on Study Experience and Learning Output.

William Schokkelé

Master Onderwijswetenschappen
Open Universiteit

Datum: 04 september 2020
Begeleiding: Dr. Stefaan Ternier

Inhoud

Samenvatting 4

Summary 5

Inleiding 6

- Aansluiting Thema. 6
- Probleemschets en Doel van het Onderzoek. 6
- Theoretisch Kader 8
  - Programmeertalen en -omgevingen. 8
  - Leren en instructie. 8
  - Feedback. 8
  - Geautomatiseerde feedback. 9
  - Invloed van geautomatiseerde feedback op leeropbrengst. 11
  - Studie-ervaring en bruikbaarheid. 11
  - Invloed van geautomatiseerde feedback op studie-ervaring. 12

Methode 14

- Ontwerp 14
- Onderzoeksgroep 15
- Materialen 16
  - De PluralSight instructievideo en slides. 16
  - Tooling experimentele groep. 16
  - Tooling controlegroep. 17
  - Toetsmateriaal: pre- en posttest. 17
  - UEQ Survey: studie-ervaring. 18
  - UEQ Benchmark. 19
- Procedure 19
  - Pretest. 20
  - Instructie. 20

Treatment. 20
Posttest. 21
Survey. 21
Analyse 21
Resultaten 23
Conclusie en discussie 27
Referenties 31
Bijlage A: de vragen van de pretest 37
Bijlage B: de vragen van de posttest 39
Bijlage C: UEQ surveyitems voor de experimentele groep 41
Bijlage D: UEQ surveyitems voor de controlegroep 42
Bijlage E: T-test – G*Power 43
Bijlage F: Cohen's d - kwalitatieve labels voor de samenhang 44
Bijlage G: UEQ benchmark en vergelijk met de experimentele groep 45
Bijlage H: resultaten T-toetsen voor vergelijk UEQ-items 46

Samenvatting

Impact van een geautomatiseerde feedbacktool op studie-ervaring en leeropbrengst

William Schokkelé

Binnen de opleiding Graduaat Informatica Programmeren stelt het team vast dat de kwaliteit van de begeleiding van studenten te lijden heeft onder de grootte van de studentengroepen (n>30): studenten zijn soms afgeleid tijdens het wachten op een lector. Soms krijgen de studenten onvoldoende hulp of feedback.

Om de studie-ervaring en leeropbrengst te verbeteren wenst het team na te gaan of bij de begeleiding van programmeeroefeningen gebruik kan gemaakt worden van een geautomatiseerde feedbacktool. Een tool die de studenten elaboratieve feedback (EF) en correctieve feedback (KCR) biedt inzetten binnen de opleiding zou de ervaren problematiek kunnen helpen oplossen.

Studenten (n = 96) Hoger Onderwijs uit de opleiding Graduaat Programmeren voeren binnen dit onderzoek een pretest uit: het bepalen van de individuele startsituatie. Hierna volgen alle deelnemers een videoles rond een nieuw leerconcept in de programmeertaal Javascript. De experimentele groep voert na de videoles een oefening uit op de nieuwe leerstof, begeleid door de geautomatiseerde feedbacktool PluralSight. De controlegroep voert deze oefening uit in een standaard onderwijssetting: begeleiding van de klasgroep door één lector. Na de oefening volgt een posttest die meet hoe goed de nieuwe leerstof werd verwerkt. Door de leeropbrengst – het verschil in score op post- en prestest - te vergelijken tussen de experimentele en de controlegroep kon nagegaan worden of het gebruik van de digitale begeleidingstool invloed heeft op de leeropbrengst. Aan de hand van de User Experience Questionnaire (UEQ) wordt de invloed op de studie-ervaring gemeten van zowel experimentele als controlegroep. De surveyresultaten voor de experimentele groep worden vergeleken met benchmarkdata met de resultaten van andere softwaretools en – voor een selectie van UEQ-items – met de resultaten van de controlegroep.

Uit dit onderzoek blijkt dat de leeropbrengst door gebruik van de digitale feedbacktool niet significant wordt beïnvloed. De UEQ-survey geeft aan dat deelnemers uit de experimentele groep (n = 63) de gebruikte tool PluralSight op 5 van de 6 schalen, aantrekkelijkheid, transparantie, efficiëntie, stimulatie en originaliteit, bovengemiddeld waarderen ten opzichte van de UEQ-benchmarkdata (n = 18483). De experimentele groep (n = 63) ondervindt een betere studie-ervaring dan de controlegroep (n = 33) op 11 van de 13 geselecteerde items van de UEQ-survey.

De specifieke setting in het programmeer onderwijs stelt ons niet in staat de bevindingen te veralgemenen, meer onderzoek is dus wenselijk.

Sleutelwoorden: geautomatiseerde feedback, leeropbrengst, studie-ervaring, hoger onderwijs

Summary

Impact of an automated feedback tool on study experience and learning output.

William Schokkelé

The team Associate's Degree Programming determines that the quality of student guidance suffers from the student's group size ($n > 30$): sometimes students are distracted while waiting for a lecturer. Sometimes students receive insufficient help or feedback.

To increase study experience and learning output the team wishes to investigate if a digital tool can be used for the guidance of programming exercises. A tool that offers elaborate feedback (EF) and corrective feedback (CF) can help solving the experienced problems.

During this research, students ($n = 96$) Higher Education of the course Associate's Degree Programming execute a pretest to determine their individual starting situation. After this, all participants receive a video lesson concerning new learning concepts in the Javascript programming language. For the experimental group the video lesson is followed by an exercise on the new learning concepts, supported by a digital learning tool. The control group is supported as in a standard course situation: guidance executed by one lecturer. This exercise is followed by a posttest measuring on which level the new learning concepts are processed. The influence of using the digital feedback tool on learning output – the difference of post- and pretest scores - can be checked by comparing the results of the experimental and control groups. The influence on study experience is measured by using the User Experience Questionnaire (UEQ) on both experimental and control group participants. The survey results of the experimental group are compared with benchmark data, containing results for other tools and – for a selection of UEQ items – with the results of the control group.

This research indicates that using the digital feedback tool has no significant influence learning output. The UEQ survey shows that participants of the experimental group ($n = 63$) evaluate the tool PluralSight as above average compared to UEQ benchmark data ($n = 18483$) on 5 out of 6 scales: attractiveness, transparency, efficiency, stimulation and originality. Participants of the experimental group ($n = 63$) experience a more positive study experience than participants of the control group ($n = 33$) on 11 items on a total of 13 of the UEQ survey.

The focus on the education in the field of computer programming restrains us from generalizing conclusions, more research is desirable.

Keywords: automated feedback, learning output, study experience, higher education

Inleiding

Aansluiting Thema.

Voor het volgen van de cursus Atelier binnen de masteropleiding Onderwijswetenschappen aan de open Universiteit Nederland sloot ik aan bij de ***Pilot Instructional Design International***. De tutors zagen een passende match met de onderzoeksgroep Computational thinking (Tactide project) onder begeleiding van dhr. Stefaan Ternier. In onderling overleg werd dit onderzoek omtrent automatisch assessment vormgegeven.

Deze studie kent ook aansluiting binnen een eerdere opdracht voor de module ***Ontwerpen van Onderwijs*** in deze masteropleiding, waarbij een 4C/ID-model werd opgesteld voor de complexe vaardigheid 'ontwikkelen van een softwareapplicatie'. De resultaten van dit onderzoek naar automatische feedback kunnen ingezet worden voor de begeleiding van deeltaakoefeningen binnen dit onderwijsmodel.

Probleemschets en Doel van het Onderzoek.

De opleiding Graduaat Programmeren is een tweejarige opleiding hoger onderwijs in dagonderwijs. De lessen worden gegeven door een team van 12 lectoren. De opleiding wordt gevolgd door 250 studenten, waarvan 150 studenten les volgen in het eerste jaar. De eerstejaarsstudenten worden opgedeeld in klasgroepen van 30 tot 35 studenten. De opleiding hanteert blended learning (Graham, 2006), waarbij onderwijsactiviteiten in contact- en afstandsonderwijs aangewend worden: hoor- en werkcolleges op campus met daarnaast opdrachten en leertaken die vanop afstand uitgevoerd worden. In deze opdrachten ontwikkelen de studenten zelfstandig of in een beperkte teams van 4 tot 5 studenten softwareapplicaties. De opdrachten worden tegen deadlines ingediend en geëvalueerd door de lectoren. Op deze wijze ontvangen de studenten voor elke module uit de opleiding 2 tot 3 keer een evaluatie en feedback alvorens de examens aan te vatten.

Binnen deze organisatie van de onderwijsactiviteiten ervaart het lectorenteam dat de studenten het soms moeilijk hebben om de gestelde leerdoelen te bereiken. De opdrachten of examens leiden al te vaak tot zwakke scores en zorgen voor een grote uitval van studenten. Basisonderdelen van de leerstof blijken vaak onvoldoende ingeoefend en verwerkt. Deze studie focust op het aspect van inoefening en verwerking van deze basisonderdelen.

Het aantal studenten in een klas heeft invloed op de kwaliteit van de leeromgeving (Ehrenberg, Brewer, Gamoran, & Willms, 2017). Hoe groter de klasgroep waaraan wordt lesgegeven hoe minder interactief contact er met de leerling is, binnen een traditionele context van frontaal klassikaal onderwijs. Toch is interactie en betrokkenheid van studenten belangrijk, zeker bij het lesgeven in grote groepen (Luscombe & Montgomery, 2016). Begeleiding van studenten tijdens klassikale oefeningen wordt moeilijker naarmate de grootte van de studentengroep stijgt (Delialioğlu, 2012; Exeter,

Ameratunga, Ratima, Morton, Dickson, Hsu, & Jackson, 2010; Mulryan-Kyne, 2010). Studenten kunnen moeilijk vragen stellen en ontvangen vaak geen begeleiding of directe feedback. Binnen de opleiding Graduaat Programmeren wordt lesgegeven aan grote groepen ($n > 30$) en is het begeleiden van de studenten dus onderhevig aan de geschetste problematiek.

Het profiel van ICT-ontwikkelaar, het afstudeerprofiel van de opleiding, bevindt zich op de Vlaamse lijst met knelpuntberoepen (VDAB, 2019), dit zijn beroepen waarvoor werkgevers het moeilijk hebben om geschikte kandidaten te vinden. De grote kans om snel een plaats te vinden op de arbeidsmarkt werkt de stijging van het aantal studenten in de hand. Daar de opleiding voor het samenstellen van het lectorteam echter ook een beroep doet op dezelfde professioneel geschoolde IT'ers, en er dus reeds een schaarste is in het profiel van potentiële nieuwe lectoren, leidt het stijgende studentenaantal niet tot een evenredige stijging in het aantal lectoren. Deze evolutie zorgt voor een verdere aanscherping van de gestelde problemen en dus een grotere relevantie van dit onderzoek.

Het lectorenteam meent dat een betere begeleiding bij het maken van oefeningen de studenten zou kunnen helpen de leerdoelen beter te bereiken. Met het oog op het verbeteren van de studie-ervaring en het behalen van betere studieresultaten is het opportuun de huidige aanpak in vraag te stellen en op zoek te gaan naar oplossingen ter verbetering. Het bijsturen van de onderwijsaanpak waarbij de studenten digitale begeleidingstools gebruiken, kan een middel zijn om dit doel te bereiken. Dit onderzoek zet in op het beter begeleiden van de studenten tijdens het inoefenen van de basisconcepten van de leerstof.

Kunnen nieuwe leerinhouden en programmeervaardigheden even goed verworven worden wanneer de studenten gebruik maken van een geautomatiseerde feedbacktool dan bij begeleiding van een grote studentengroep ($n>30$) door één lector? Wat is het effect van gebruik van deze tool op de leeropbrengst en de studie-ervaring? Hoe verhoudt de gebruikerservaring met de feedbacktool zich tot de gebruikerservaring met ander tools?

Dit onderzoek gaat na of er een relatie is tussen het aanbieden van geautomatiseerde feedback op de leeropbrengst en de studie-ervaring van studenten Hoger Onderwijs, met als doel de studenten een betere begeleiding te kunnen geven dan op heden het geval is. Er wordt onderzocht of het geven van geautomatiseerde feedback bij het maken van oefeningen dit gebrek aan begeleiding kan verhelpen. Er wordt nagegaan of studenten die gebruik maakten van een automatische feedbacktool minstens even goed presteren op een toets dan studenten die les kregen in de standaard onderwijscontext, en hoe de studenten de geautomatiseerde begeleiding ervaren in vergelijking met de begeleiding door één lector. De gebruikerservaring met de tool PluralSight wordt vergeleken met benchmarkdata van andere tools.

Theoretisch Kader

Programmeertalen en -omgevingen.

Voor het ontwikkelen van software wordt gebruik gemaakt van programmeertalen. De Graduaatsopleiding Informatica Programmeren, waar dit onderzoek wordt uitgevoerd, maakt gebruik van een waaier aan verschillende programmeertalen en -technologieën. Web Development is een van de pijlers van het opleidingsprogramma is. De studenten gaan binnen dit opleidingsdomein aan de slag met de programmeertaal Javascript. Javascript is momenteel een van de meest gebruikte programmeertalen wereldwijd (Altherwi & Gravell, 2019; Theisen, 2019). De keuze om toe te spitsen op opdrachten in deze programmeertaal draagt dan ook bij aan de relevantie van dit onderzoek.

Leren en instructie.

Leren wordt omschreven als een blijvende verandering in prestatie of prestatievermogen als gevolg van ervaring en interactie met de wereld (Driscoll, 2014). Het leren van een programmeertaal vereist het verwerven van de correcte syntax, semantiek en logische denkpatronen. Dergelijke kennisverwerving gebeurt niet automatisch door de lerende, maar is, zoals aangegeven in de cognitive-load theorie (Sweller, 2019), het resultaat van een volgehouden inspanning van de lerende, bijgestaan door expliciete instructie. Die instructie kan face-to-face plaatsvinden of – zoals in dit onderzoek – met digitale tools.

Doordat de opslagcapaciteit van het kortetermijngeheugen beperkt is tot een 7-tal informatie-eenheden is het belangrijk leerstof in te bedden in het langetermijngeheugen (Sweller, 2019). Nieuwe leerinhouden moeten zoveel mogelijk worden gekoppeld aan eerder verworven kennis en vaardigheden. Het gebruik van een advance organizer kan het leren dus positief beïnvloeden (Valcke, 2017). Het beschikken over voldoende achtergrondkennis, over informatie die in jouw brein zit, kan je helpen in de manier waarop je omgaat met nieuwe gegevens (Heusser, Awipi, & Davachi, 2013).

Van zodra we leren start echter ook het proces van vergeten: de nieuwe leerstof gaat na verloop van tijd verloren. Door oefening en herhaling kunnen we leerstof sneller opfrissen en langer beschikbaar houden in het langetermijngeheugen (Hamel, Côté, Matte, Lepage, & Bernier, 2019; Heusser et al., 2013). Het herhaaldelijk inoefenen van nieuwe vaardigheden zorgt voor transfer naar het langetermijngeheugen.

Feedback.

Feedback wordt omschreven als informatie gegeven door een verstrekker (b.v. leraar, gelijke, boek, ouder, zichzelf, ervaring) over iemands prestatie of begrip (Hattie & Timperley, 2007). Wanneer het geven van feedback de mogelijkheid schept tot correctie, worden feedback en instructie verweven: het

feedbackproces zelf neemt de vorm aan van nieuwe instructie (Kulhavy, 1977). Zo kunnen studenten feedback gebruiken om het leerproces te verbeteren, het inzicht aan te scherpen.

Feedback kan inhoudelijk opgedeeld worden in drie categorieën (Shute, 2008): kennis van resultaat (KR), kennis van correct resultaat (KCR) en elaboratieve feedback (EF). KR geeft aan of het gegeven antwoord goed of fout is, terwijl KCR ook het goede antwoord laat zien. Bij EF krijgt de student aanvullende informatie over het antwoord of over de lestopic, worden specifieke fouten toegelicht of voorbeeldуitwerking en begeleiding gegeven. Uit meta-analyse van de resultaten van 40 studies blijkt dat EF, het leveren van een uitleg als feedback, effectiever is dan KR en KCR (Van der Kleij, Feskens, & Eggen, 2015). Naast de inhoud van de feedback speelt ook timing van de feedback een rol (Price, Handley, Milar, & O'Donovan, 2010; Sadler, 2010; Shute, 2008). Feedback kan onmiddellijk gegeven worden of uitgesteld.

Door de verschillende facetten zoals inhoud en timing waaruit feedback bestaat is het geven van effectieve feedback gecompliceerd (Agricola, Prins, & Sluijsmans, 2020; Winstone & Carless, 2019): de relatie tussen vorm, timing en effectiviteit van feedback is complex en veranderlijk (Price et al., 2010; Sadler, 2010). Stobart (2008) geeft aan dat drie condities vervuld moeten zijn opdat feedback effectief en nuttig zou zijn: 1) de lerende heeft de feedback nodig, 2) de lerende heeft tijd om de feedback te gebruiken en 3) de lerende kan de feedback gebruiken.

Geautomatiseerde feedback.

Terwijl feedback gegeven door een echte tutor heel wat subtiele gradaties kent, zoals het aanduiden van een fout, het geven van een tip, het aanbieden van een (alternatieve) oplossing en zelfs het geven van een schouderklopje, is geautomatiseerde feedback beperkter. In een maatschappij met voortschrijdende inzichten in vakgebieden zoals Artificial Intelligence zullen tools ontwikkeld worden met verregaande mogelijkheden op het gebied van communicatie met gebruikers en het geven van feedback. Het onderhavige onderzoek richt zich niet op de ontwikkeling van nieuwe tools, maar op mogelijkheden om bestaande tools te integreren binnen de huidige opleiding.

Bij het geven van feedback zijn digitale annotaties en opnames goed bruikbaar voor studenten. (Ryan et al., 2019). Het gebruik van een geautomatiseerde feedbacktool kan zo misschien een antwoord bieden op de verzuchting van veel studenten die aangeven dat feedback gegeven door lectoren moeilijk te begrijpen is en dat het tot leidt tot frustratie en ontgoocheling wanneer de gegeven feedback onduidelijk, te bondig of niet behulpzaam is (Ferguson, 2011; Hounsell, D., McCune, Hounsell, J., & Litjens, 2008; Hyland, 2013).

In het domein van programmeren werden tools ontwikkeld die geautomatiseerde feedback bieden bij het maken van oefeningen. Enkele bestaande tools werden vergeleken voor dit onderzoek: FreeCodeCamp, Scrimba, CodeAcademy, CodeInGame, HackerRank, Dodona, Learncs, Sphere

Online Judge en PluralSight. Tabel 1 toont hoe deze platformen voor programmeeroefeningen scoren volgens volgende criteria:

(1) Technologie: is de tool geschikte voor de gekozen programmeertaal Javascript?
(2) Videolessen: biedt de tool de mogelijkheid tot instructie met behulp van videolessen?
(3) Licentie: is de tool geschikt voor gebruik door de studentengroep, zijn er geen belemmeringen door extra aan te schaffen licenties?
(4) In welke mate geeft de tool feedback: KR, KCR, EF.

Tabel 1

De mogelijkheden van de verschillende tools voor automatische feedback, getoetst aan de selectiecriteria.

Tool	Technologie	Videolessen	licentie	KR	KCR	EF
CodeAcademy	x	x		x	x	
CodeInGame	x		x	x	x	x
Dodona	x		x	x	x	
FreeCodeCamp	x		x	x	x	x
HackerRank	x		x	x	x	
Learncs			x	x	x	x
PluralSight	x	x	x	x	x	x
Scrimba	x	x		x	x	
Sphere Online Judge	x			x	x	

Op basis van deze criteria werd de tool PluralSight geselecteerd voor dit onderzoek. De tool PluralSight combineert de mogelijkheid om videolessen met slides in te zetten voor instructie en interactieve oefeningen met geautomatiseerde correctieve en elaboratieve feedback. PluralSight is de enige tool die beantwoordt aan alle vooropgestelde criteria.

Slechts 3 tools bieden instructie aan via videolessen: CodeAcademy, PluralSight en Scrimba. De online oefeningeneditoren van Scrimba en CodeAcademy voorzien echter enkel in feedback van het type KR en KCR. Scrimba en CodeAcademy zijn betalende tools. Deze tools inzetten zou dus voor de studenten een meerkost betekenen. Echter, ook aan het gebruik van PluralSight is een abonnementskost verbonden, maar die is reeds vervat in het inschrijvingsgeld voor de opleiding waar dit onderzoek plaatsvond. Een meevaller voor dit onderzoek, maar een aandachtspunt voor eventueel vervolgonderzoek.

PluralSight biedt de lerende onmiddellijke feedback. De tool kan feedback geven zowel op syntactisch als op semantisch vlak: naast het controleren op spellingsfouten kan de tool nagaan of de

deelnemer de correcte programmastructuur aanwendt. Tijdens het maken van oefeningen kunnen de leerbronnen - een lesvideo en de instructieslides - geraadpleegd worden. De lerende wordt geïnformeerd of de ingediende oplossing goed of fout is (KR), kan een voorbeeldoplossing raadplegen (KCR). De tool zet de lerende via tips op het juiste pad wanneer een fout wordt gemaakt (EF).

Dat PluralSight een wereldwijd platform is, gebruikt door ruim 700.000 developers, ontwikkeld door een professioneel team, leidt tot positieve verwachtingen omtrent de gebruikerservaring in vergelijking tot andere tools, ook binnen andere vakgebieden.

Invloed van geautomatiseerde feedback op leeropbrengst.

Hattie (2008) geeft in meta-analytisch onderzoek aan dat het geven van feedback één van de krachtige mechanismen om leerprestaties te verbeteren. De types feedback die de sterkste impact op de leerprestaties hebben zijn volgens Hattie en Timperley (2007) het geven van cues, informatie geven over de prestaties, bekrachtigen, video of audio feedback, computergestuurde feedback, de leerdoelen aanhalen bij feedback en evaluatieve feedback door de studenten.

Een digitale feedbacktool biedt voordelen ten opzichte van een standaard klassetting. De studenten krijgen onmiddellijk elaboratieve feedback, kunnen deze direct verwerken in de eigen oplossing. In een klassetting moet vaker gewacht worden, zodat sommige studenten met problemen kunnen kampen die het verder werken belemmeren. Onmiddellijk beschikbare, schriftelijke feedback verhoogt de kwaliteit van ingediende werkstukken en de evaluatiescores (Kenny & Pahl, 2009; Stevenson & Phakiti, 2014). De gekozen tool PluralSight biedt de feedback stapsgewijs aan, wat tot betere resultaten leidt dan het aanbieden van alle feedback in een keer (Shute, 2008). Deze argumenten leiden er toe dat kan aangenomen worden dat de leeropbrengst bij het gebruik van de feedbacktool minstens even goed is als in een standaard klassetting.

Studie-ervaring en bruikbaarheid.

Het gebruik van een geautomatiseerde feedbacktool verandert de studie-ervaring voor de studenten: de begeleiding door een lector wordt ingeruild voor een digitaal begeleidingsinstrument. Naast het meten van de uitkomstcijfers bij een opdracht gaat dit onderzoek na welke de implicaties van het gebruik van de geautomatiseerde feedbacktool zijn op het gebied van gebruikerservaring.

De bruikbaarheid van een softwareapplicatie is de mogelijkheid om de softwareapplicatie te begrijpen, te leren en te gebruiken in een specifieke context (Abran, Khelifi, Suryb, & Seffah, 2003). Volgens de Internationale Organisatie voor Standaardisatie (ISO, 2016) kent de bruikbaarheid van een softwareapplicatie drie criteria:

(1) Effectiviteit: de nauwkeurigheid en volledigheid waarmee gebruikers specifieke doelen kunnen bereiken

(2) Efficiëntie: de verbruikte bronnen in relatie tot de nauwkeurigheid en volledigheid
(3) Tevredenheid: het comfort en aanvaardbaarheid van het product voor de gebruikers

Voor het meten van de bruikbaarheid zijn verschillende meetinstrumenten beschikbaar. De Computer System Usability Questionnaire (CSUQ; Lewis, 1995), de System Usability Scale (SUS; Jordan, Thomas, McClelland, & Weerdmeester, 1996) en de User Experience Questionnaire (UEQ; Schrepp, 2015) werden vergeleken.

De UEQ survey drukt het construct bruikbaarheid uit in 6 goed omschreven schalen, die ook de ISO-criteria omvatten. De lijst van items is beschikbaar in de Nederlandse taal en de onderzoekers hebben toegang tot benchmarkdata met resultaten van eerdere studies. Deze argumenten leidden tot de keuze voor deze UEQ-survey. De originele Duitse versie van UEQ werd ontwikkeld in 2005, waarbij met behulp van data-analyse technieken de praktische relevantie van de gebruikte schalen werd verzekerd (Schrepp, 2015). De UEQ vragenlijst werd op heden vertaald in 21 talen waaronder het Nederlands en levert het beste resultaat indien toegepast onmiddellijk na een interventie (Schrepp, 2015). Met behulp van een meegeleverde Excel-tool kan een data-analyse van de resultaten bekomen worden. Benchmarkdata is meegeleverd over de schaalgemiddelden van 18483 personen uit 401 verschillende studies (Schrepp, 2015). Deze schaalgemiddelden stellen ons in staat de ervaringen met de gekozen feedbacktool te vergelijken met een grote onderzoekspopulatie, waardoor conclusies over de relatieve kwaliteit ten opzichte van andere tools kunnen geopperd worden.

De studie-ervaring met een digitale begeleidingstool wordt met behulp van de UEQ nagemeten op 6 categorieën:

(1) Aantrekkelijkheid: de algemene indruk van de tool.
(2) Transparantie: de mate waarin het gemakkelijk is om vertrouwd te raken met de tool en die te leren gebruiken
(3) Efficiëntie: de mate waarin de respondenten hun taken kunnen uitvoeren zonder onnodige inspanning
(4) Bestuurbaarheid: de mate waarin de deelnemer zich in controle van de interactie voelt, de interactie als betrouwbaar wordt aangegeven.
(5) Stimulatie: de mate waarin het motiverend en leuk gevonden wordt om de tool te gebruiken
(6) Originaliteit: de mate waarin het ontwerp creatief is, de interesse van de deelnemer wekt.

Invloed van geautomatiseerde feedback op studie-ervaring.

Geautomatiseerde schriftelijke feedback verhoogt de kwaliteit van ingediende schrijfopdrachten (Stevenson & Phakiti, 2014), verhoogt de evaluatiescore (Kenny & Pahl, 2009) en leidt tot goede scores voor welbevinden (Kenny & Pahl, 2009). Een computer gebaseerde tool lost de discrepantie

tussen de huidige status van de student en bedoelde leeruitkomsten beter op dan een traditionele onderwijsomgeving (John Hattie & Gan, 2011). Hierbij is stapsgewijze feedback effectiever dan het aanbieden van alle feedback in een keer (Shute, 2008). Kort geformuleerde wenken en tips zijn effectiever dan exhaustief geformuleerde feedback (Lee, 2019).

Geautomatiseerde regels zijn niet afdoende om studenten met grote conceptuele problemen te helpen (Singh, Gulwani, & Solar-Lezama, 2012). Wanneer feedback impliceert dat de oplossing van de student zo goed als volledig herschreven moet worden resulteert deze feedback dus eerder in een open vraag.Vraagstellingen en hypothesen.

De centrale vraag binnen dit onderzoek luidt: “Wat is de impact van het gebruik van een geautomatiseerde feedbacktool op de leeropbrengst en studie-ervaring van startende studenten Hoger Onderwijs, binnen een opleiding programmeren?”. We onderscheiden de volgende deelvragen:

(1) Is de leeropbrengst gemeten bij het gebruik van de geautomatiseerde feedbacktool minstens even groot als bij feedback door de lector in de klas?
(2) In welke mate ervaren de studenten de feedbacktool als aantrekkelijk, transparant, efficiënt, bestuurbaar, stimulatief en origineel, in vergelijking met andere tools?
(3) Wordt geautomatiseerde feedback even goed ervaren als de feedback gegeven door een lector?

De redeneringen opgebouwd in het theoretisch kader leiden hierbij tot de volgende hypothesen:

H1: Startende studenten in een programmeeropleiding Hoger Onderwijs behalen bij het gebruik van de geautomatiseerde feedbacktool PluralSight een minstens even hoge leeropbrengst dan studenten in een traditionele onderwijssetting.
H2: Startende studenten in een programmeeropleiding Hoger Onderwijs waarderen de tool PluralSight in elk van de 6 UEQ-schalen even goed of beter dan de tools uit de Benchmarkresultaten.
H3: Startende studenten in een programmeeropleiding Hoger Onderwijs waarderen geautomatiseerde feedback door de tool PluralSight minstens even goed als de feedback gegeven door een lector in een traditionele onderwijssetting.

Methode

Ontwerp

Dit kwantitatief onderzoek presenteert een studie met een quasi-experimenteel pretest-posttest design. De keuze voor een quasi-experimenteel ontwerp volgt uit de onderverdeling van de participanten bij de start van het academiejaar in klasgroepen. De experimentele groep wordt gevormd door 2 klasgroepen, de controlegroep door 1 klasgroep. Pretest-posttest design werd aangewend om voor de participanten vast te kunnen stellen wat de studieprestatie is voor en na het experiment en de leeropbrengst als verschil in deze scores te kunnen berekenen.

Het gebruik van een pretest op eerder geleerde programmeerconcepten, zoals in dit onderzoek toegepast, kan het leren positief beïnvloeden als advance organizer (Valcke, 2017). Tijdens de pretest, het treatment en de posttest maken de studenten binnen dit onderzoek geëvalueerde opdrachten die dit leerproces positief ondersteunen (Stobart, 2008). Dit in tegenstelling tot de opvatting dat toetsen enkel bedoeld zijn voor summatieve doelen, welke aanleiding kan geven tot *teaching to the test* (Birenbaum et al., 2006).

De studenten zullen tijdens de instructiefase niet enkel passief luisteren maar actief deelnemen door oefeningen te maken. Door gebruik te maken van videolessen en slides is het mogelijk om de lessen aan te bieden aan verschillende studentengroepen en zo de variatie in aanpak en stijl van de lector uit te schakelen: elke studentengroep krijgt dezelfde videoles aangeboden.

In dit onderzoek zullen voor de treatmentfase zowel de experimentele groep als de controlegroep feedback genieten. De digitale tool levert direct feedback op een antwoord (onmiddellijke EF) en staat toe de voorbeeldoplossing zichtbaar te maken (KCR). In de standaardsetting kan de student de lector raadplegen. Deze levert de directe EF. Stobart (2008) gaf aan dat drie condities vervuld moeten zijn opdat feedback effectief en nuttig zou zijn: 1) de lerende heeft de feedback nodig, 2) de lerende heeft tijd om de feedback te gebruiken en 3) de lerende kan de feedback gebruiken. De eerste conditie is vervuld wanneer er een kloof is tussen de huidige kennis en het leerdoel (John Hattie & Timperley, 2007). In dit onderzoek brengen we een nieuw leerstofonderdeel aan en beantwoorden dus aan dit criterium. Om voor zoveel mogelijke deelnemers te voldoen aan de tweede conditie wordt het experiment uitgevoerd in een ruim bemeten tijdskader, de beschikbare tijd is voor de experimentele groep en de controlegroep gelijk. Of de deelnemer werkelijk voldoende tijd heeft om de ontvangen feedback te gebruiken hangt weliswaar af van de individuele aanleg en inzet. Bij de controlegroep zal het verkrijgen van feedback in een standaard klassituatie afhangen van het initiatief van de lerende om een vraag te stellen en de mogelijkheid van de lector zich beschikbaar te stellen. Voor het vervullen van de derde conditie wordt gezorgd voor het gebruik van de correcte leertools zodat de feedback geïmplementeerd kan worden.

Dit onderzoek spitst zich toe op strak omlijnde competenties, basisconcepten van het programmeren. Zien we het ontwikkelen van een softwareapplicatie als een complexe vaardigheid binnen een onderwijsmodel zoals 4C/ID (Van Merriënboer & Kirschner, 2013), dan richt dit onderzoek zich dus op deeltaakoefeningen voor het inoefenen van specifieke vaardigheidsaspecten namelijk het gebruik van for- en while-lussen als bouwsteen voor het programmeren. Doordat dit onderzoek focust op het aanleren van basisconcepten van het programmeren sluiten we grotere conceptuele problemen, waar automatische feedback moeilijker hulp kan bieden (Singh et al., 2012), uit.

Voor het beantwoorden van de onderzoeksvragen worden twee verschillende evaluatiemethoden aangewend (Kinshuk & Russell, 2000), de scores van een programmeeropdracht en de antwoorden op een vragenlijst. Deelvraag 1 vergelijkt de studieprestatie bij gebruik van de geautomatiseerde feedbacktool met de studieprestatie in een standaard onderwijssetting: één lector die een grote groep studenten begeleidt (n>30). De studieprestatie wordt nagegaan door het maken van een programmeeropdracht. Het antwoord van de deelnemers wordt afgetoetst ten opzichte van een verbetersleutel. Om de onderzoeksvragen 2 en 3 te beantwoorden nemen de respondenten deel aan een survey. Hiertoe maakt het onderzoek gebruik van de UEQ vragenlijst (Schrepp, 2015).

Onderzoeksgroep

De deelnemers aan het onderzoek zijn eerstejaarsstudenten aan de opleiding Graduaat Informatica Programmeren aan een Vlaamse Hogeschool. De volledig opleiding heeft een duur van 2 academiejaren. De opleiding kent 150 eerstejaarsstudenten waarvan 120 les volgen in het dagtraject en 96 studenten deelnamen aan het onderzoek. De studenten die startten in het avond- of weekendtraject (n = 30) werden niet in het onderzoek betrokken, daar de persoonsgebonden variabelen van deze studiegroep te veel afwijken van de daggroepen: deze studenten hebben een hogere gemiddelde leeftijd, combineren vaak studie met werk en hebben dus ook een andere motivatie dan de generatiestudenten in de dagtrajecten. De studenten hebben allen minimaal de leeftijd van 18 jaar.

Binnen de Graduaatsopleiding worden de studenten aan het begin van het academiejaar ingedeeld in klasgroepen met een groepsgrootte tussen de 30 en 35 studenten. Deze indeling gebeurt door de dienst studentenadministratie van de school en is gebaseerd op de inschrijvingsdatum. De 3 studiegroepen volgen les op andere momenten, bij verschillende lectoren. 2 Klasgroepen fungeren als experimentele groepen (n = 63) en één klasgroep als controlegroep (n = 33). Wanneer je elke klasgroep zou splitsen in een experimentele en een controlegroep valt de grootte van de controlegroep in de standaardsetting terug op 15 studenten voor een lector, wat niet representatief is voor dit onderzoek, waarbij een groepsgrootte van 30 werd verondersteld. Feedback geven aan 15 studenten is anders dan feedback geven aan 30 studenten. Dat de lessen voor de klasgroepen apart doorgaan en

vaak gegeven worden door verschillende lectoren verklaart de keuze voor videolessen voor het instructiegedeelte van het onderzoek.

Dit onderzoek bestaat uit een pretest, een treatment, een posttest en een survey. Om deze items gekoppeld te houden kreeg elke deelnemer een unieke code toegewezen, bestaande uit 8 willekeurige letters en cijfers.

Materialen

De PluralSight instructievideo en slides.

De deelnemers van zowel experimentele groep als de controlegroep krijgen een 15 minuten durende instructievideo te zien via het online platform PluralSight over het topic ***JavaScript: Using While and For Loops***. Naast het videomateriaal zijn ook de instructieslides ter beschikking.

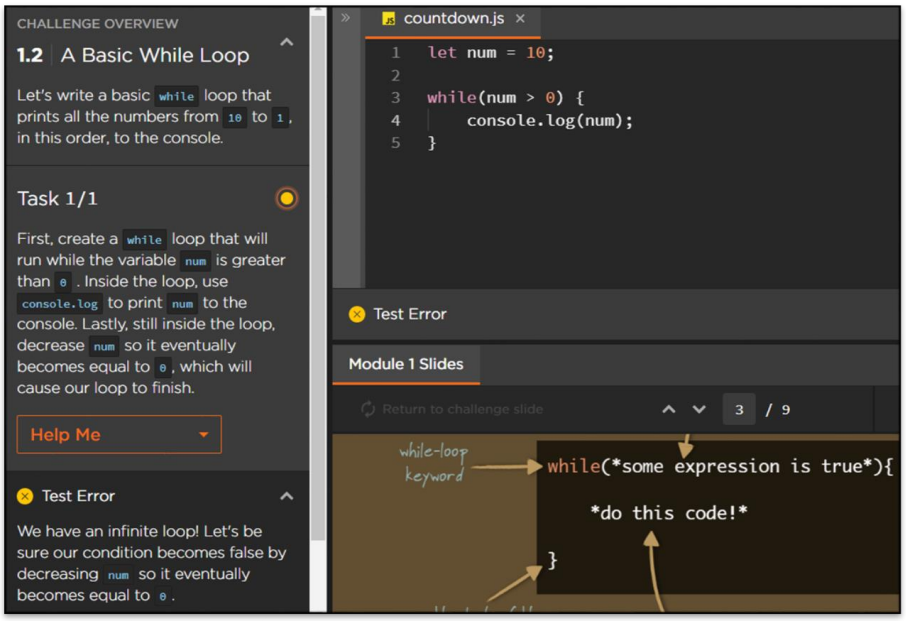
Tooling experimentele groep.

De online tool PluralSight werd ingezet om de experimentele groep te voorzien van geautomatiseerde feedback bij het maken van oefeningen rond basisconcepten programmeren.

Figuur 1 toont de gebruikersinterface van de tool PluralSight. De tool formuleert de opgave in volzinnen. Indien van toepassing wordt een omvangrijke opgave opgesplitst in deeltaken. De lerende kan de oefening maken, een oplossing indienen en ontvangt feedback:

(1) Kennis van resultaat (KR): de lerende weet of de eigen oplossing correct of fout is.
(2) Correctieve feedback (onmiddellijke KCR): de voorbeeldoplossing is beschikbaar via het onderdeel ***Help me***.
(3) Elaboratieve feedback (onmiddellijke EF): een fout wordt aangegeven met hierbij instructie hoe de oplossing kan verbeterd worden. Hiernaast worden de instructieslides continu terbeschikking gesteld, het videomateriaal is raadpleegbaar via het onderdeel ***Help me***.

Voorbeelden van gegeven automatische feedback door de tool PluralSight zijn: ‘Let's use the while keyword to create a while loop’ en ‘We have an infinite loop! Let's be sure our condition becomes false by decreasing num so it eventually becomes equal to 0’.



Figuur 1. Interactieve oefening in de tool PluralSight. Bovenaan rechts staat het opdrachtgedeelte. Hier maakt de deelnemer de oefening door programmeercode in te tikken. Daaronder krijgt de deelnemer toegang tot de instructieslides. Bovenaan links staat de formulering van de algemene opdracht. Daaronder een beschrijving van de huidige deeltaak. Onderaan links wordt de feedback getoond. De feedback geeft aan of de oplossing juist of fout is en geeft bij een incorrect antwoord aanwijzingen om tot een goede oplossing te komen.

Tooling controlegroep.

De deelnemers uit de controlegroep gebruiken voor het maken van de oefening de standaardtool uit de lessen: de Integrated Development Environment (IDE) Visual Studio Code. Deze ontwikkeltool biedt de studenten autocompletion of het automatisch aanvullen van codewoorden en duidt syntaxfouten aan. Visual Studio Code geeft niet aan of de oplossing juist of fout is: als een conceptueel foutieve oplossing geen syntaxfouten bevat, dan zal deze toch uitgevoerd worden. De student wordt niet op de hoogte gesteld van het probleem. Visual Studio Code levert dus geen KR, geen EF of KCR, wat wel aanwezig is bij PluralSight.

Toetsmateriaal: pre- en posttest.

De eerste deelvraag wordt getoetst door middel van een kwantitatieve registratie en analyse van toetsresultaten. De variabele leeropbrengst wordt gekwantificeerd door het scoreverschil tussen pre- en

posttest te berekenen. Dit gebeurt op een identieke manier voor de experimentele als de controlegroep: de testvragen, de gewichten per vraag en de evaluatie ervan zijn voor beide onderzoekgroepen gelijk.

Zowel pre- als posttest bestaan uit multiple choice vragen. De vragen worden beantwoord via een digitaal formulier. Bijlage A bevat de vragen voor de pretest, bijlage B bevat de vragen voor de posttest. Aan elk van deze vragen werd een gewicht gekoppeld met waardes van 1 tot 5. De maximale score voor zowel de pre- als de posttest is 15. Het gewicht werd per vraag bepaald aan de hand van de complexiteit van de vraag, waarbij het laagste gewicht duidt op de minst complexe vraag. Vragen die enkel testen op het syntactisch correct zijn van een enkelvoudig programmeerstatement kregen een gewicht 1, vragen die meerdere programmeerstatements bevatten, waarbij verschillende programmeerconcepten zoals conditionele- en lusstructuren werden genest kregen een gewicht 5.

De pretest behelst vragen over de reeds gekende leerstof tot op heden, de posttest bevat vragen over de nieuwe leerstof uit het onderzoek. De studenten krijgen het resultaat van deze tests steeds op het einde van de test (uitgestelde KR) te zien.

UEQ Survey: studie-ervaring.

Voor het meten van de studie-ervaring beantwoordden de deelnemers de User Experience Questionnaire UEQ (Laugwitz et al., 2008; Schrepp, 2015; Schrepp et al., 2014). De originele Duitse versie van UEQ werd ontwikkeld in 2005, waarbij met behulp van data-analyse technieken de praktische relevantie van de gebruikte schalen werd verzekerd (Schrepp, 2015). De UEQ vragenlijst werd op heden vertaald in 21 talen waaronder het Nederlands en levert het beste resultaat indien toegepast onmiddellijk na een interventie (Schrepp, 2015).

De deelnemers uit de experimentele groep beantwoordden de 26 items die samen de 6 schalen van de survey uitmaken. De consistentie, aangegeven door de waarde voor Cronbach's alpha, van de UEQ schalen aantrekkelijkheid (n = 6, α = .81), transparantie (n = 4, α = .80), efficiëntie (n = 4, α = .78), bestuurbaarheid (n = 4, α = .56), stimulatie (n = 4, α = .79) en originaliteit (n = 4, α = .70) en de validiteit van de UEQ schalen werd getoetst in 401 studies (n = 18483) en bruikbaar bevonden (Schrepp, 2015). Deelname van 20-30 personen geeft reeds stabiele onderzoeksresultaten (Schrepp, 2015). De surveyvragen werden beantwoord op een semantisch differentiaal-schaal (Osgood, Suci, & Tannenbaum, 1957) met 7 punten (Schrepp, 2015). Voorbeelden hiervan zijn de tegengestelde begrippen onbegrijpelijk en begrijpelijk, langzaam en snel, belemmerend en ondersteunend, verwarrend en overzichtelijk. De helft van de items start met positieve term, de andere helft met de negatieve (Schrepp et al., 2017). Het was dan ook noodzakelijk de data-items na de survey te transformeren in één richting. Bijlage C geeft een overzicht van alle items met telkens de tegengestelde begrippen van de semantisch differentiaalschaal en de schalen waartoe de items behoren.

Om de impact op de studie-ervaring van de feedbacktool te vergelijken met de controlegroep, die feedback door een lector ontvangt in een reguliere klassituatie, is het nodig vergelijkingspunten te vinden tussen de experimentele groep en de contolegroep. 13 Items uit de UEQ waardenschaal werden gekozen om de ontvangen feedback van een lector te enquêteren. De tegenstellingen tussen onbegrijpelijk tot begrijpelijk, complex tot eenvoudig, overzichtelijk tot verwarrend kunnen afgetoetst worden op feedback door een lector, terwijl dit voor de tegenstellingen zoals afstotend tot aantrekkelijk, onpraktisch tot praktisch niet of minder het geval is. Van een tool is het bijvoorbeeld van belang of deze aantrekkelijk en overzichtelijk werd vormgegeven: de schikking van elementen zoals knoppen en keuzelijsten, het gebruik van kleuren en contrast. Echter, het oordeel of een lector al of niet aantrekkelijk wordt bevonden door de studenten, draagt niet bij tot de kwaliteit van het onderzoek. Sommige items uit de UEQ-waardenschaal - die toepasbaar zijn op softwaretools - zijn dus voor een menselijke feedbackgever ongeschikt om te beoordelen. Deze items werden niet weerhouden. De geselecteerde items worden opgelijst in bijlage D.

UEQ Benchmark.

De UEQ-survey bevat een benchmarktool (Schrepp, Hinderks, & Thomaschewski, 2017) die toestaat de onderzoeksresultaten te vergelijken met de data van 18483 personen uit 401 eerdere studies. Waar slecht één enkele UEQ-meting bestaat, is het moeilijk de kwaliteit van een product in te schatten. Wanneer de resultaten kunnen vergeleken worden met andere tools, kan beter ingeschat worden hoe de tool gewaardeerd wordt. Deze tool is in Excel-formaat online beschikbaar.

De benchmarktool bevat voor elk van de UEQ-schalen de berekende interne consistentie a.d.h.v. Cronbach's alpha en de numerieke grenswaarden voor volgende 5 categorieën (Schrepp et al., 2017):

(1) Excellent: het geëvalueerde product scoort bij de 10% hoogste resultaten
(2) Goed: 10% scoort hoger, 75% scoort minder dan het geëvalueerde product
(3) Boven gemiddeld: 25% scoort hoger, 50% scoort minder dan het geëvalueerde product
(4) Onder gemiddeld: 50% scoort hoger, 20% scoort minder dan het geëvalueerde product
(5) Slecht: het geëvalueerde product is bij de 25% laagste resultaten

Procedure

Het onderzoeksvoorstel werd door het Research Ethics Commitee (cETO) goedgekeurd op de aspecten van de Algemene Verordening Gegevensbescherming (AVG), en de manier waarop proefpersonen in het onderzoek worden betrokken (o.a. informed consent, informatiebrieven). De respondenten werden door de onderzoeker uitgenodigd om vrijwillig deel te nemen aan het onderzoek en gaven hun toestemming tot deelname.

Als voorbereiding op het onderzoek werden via de online tool Random Code Generator (Digital Devils, 2020) voor elk van de deelnemers (*n*=96) een unieke code gegenereerd bestaande uit 8 willekeurige letters en cijfers. Deze codes werden op afzonderlijke labels afgedrukt.

Het onderzoek vond plaats in de geselecteerde klasgroepen tijdens een les die deel uitmaakte van het regulier lestraject. Het onderzoek startte met het verdelen van de willekeurige codes, waarbij elke student een codelabel trok uit een trekzak. Deze code werd door de deelnemer op de digitale formulieren voor pre- en posttest en de UEQ-surveyvragen ingevuld. Op deze manier kan de onderzoeker enkel de koppeling tussen presest, posttest en survey maken, een terugkoppeling naar de individuele student is onmogelijk.

De opeenvolgende stappen van het onderzoek en de dataverzameling bestaan uit de pretest, de instructiefase, de treatmentfase, de posttest en de survey. Deze verschillende fases worden in chronologische volgorde toegelicht.

Pretest.

Alle deelnemers uit zowel experimentele als controlegroep voeren de pretest uit in dezelfde setting: in de klas, individueel, zonder hulp van tools of de lector. Onmiddellijk na de test ziet de deelnemer de score (KR). Bijlage A toont de vragen van de pretest.

Instructie.

Alle deelnemers volgen via het platform PluralSight een interactieve videoles van 15 minuten over een nieuw lesonderwerp: ***JavaScript: Using While and For Loops***. Naast de video met instructie kunnen de deelnemers slides raadplegen waarop de leerstof wordt uitgelegd. De leerstof die in dit onderzoek aangereikt wordt kadert in het opleidingstraject van de opleiding Graduaat Programmeren. De leerinhoud van dit topic wordt ingebed in eerder verworven kennis en vaardigheden: variabelen, operatoren en bewerkingen.

De keuze voor deze interactieve videoles sluit uit dat er een lector specifieke invloed is over de verschillende groepen waarbij de mate van instructie kan variëren of ingegaan kan worden op specifieke problemen.

Treatment.

Alle deelnemers voeren dezelfde oefening uit op deze nieuwe leerstof. Deze onderzoeksfase bestaat uit 5 programmeeroefeningen op het nieuwe lesonderwerp. De experimentele groep voert deze challenges uit via het platform PluralSight en ontvangt dus geautomatiseerde feedback. De controlegroep voert deze oefeningen uit met behulp van de standaardtool uit de lessen de Integrated Development Environment (IDE) Visual Studio Code. De controlegroep wordt begeleid door één lector die de studenten individuele feedback geeft. In de controlegroep wordt dus door de lector niet

klassikaal ingegaan op problemen, dit zou kunnen leiden tot een groter leereffect vanwege het leren van medestudenten die vragen stellen. De feedback (EF) in de controlegroep bleef beperkt tot het geven van een hint of het wijzen op een denkfout, er wordt geen theorie opgefrist of extra lesgegeven. De studenten in de experimentele en de controlegroep hebben een voorbeeldoplossing ter beschikking (onmiddellijke KCR).

Door studenten leerstof op een actieve manier eigen te laten maken en hen (al dan niet digitaal) begeleid oefeningen te laten maken op een beperkt onderdeel van de leerstof houdt dit onderzoek rekening met de bevindingen omtrent het leerproces, de aangegeven moeilijkheden door cognitieve overbelasting en de bevindingen omtrent het vergeetproces.

Posttest.

Alle deelnemers voeren de posttest uit. Alle respondenten maken deze posttestoefening in dezelfde setting: geen digitale tools of hulp van de lector. Onmiddellijk na de test ziet de deelnemer de score (KR). Bijlage B toont de vragen van de posttest.Uit de scores voor pre- en posttest kan de leeropbrengst worden berekend en kan een antwoord geformuleerd worden op de eerste deelvraag.

Survey.

Als onderdeel van het experimenteel ontwerp wordt een survey afgenomen bij studenten uit beide groepen. In de experimentele groep werd de volledige UEQ-survey afgenomen om gebruikservaring te meten. Deze antwoorden kunnen vergeleken worden met de antwoorden uit de aangeleverde benchmark en geven antwoord op deelvraag 2. De controlegroep werd bevraagd met een subset van de UEQ-survey, namelijk alleen die vragen die betrekking hebben op begeleiding of feedback van een lector. Vergelijking van de antwoorden op deze subset tussen de experimentele groep en de controlegroep geeft antwoord op deelvraag 3.

Analyse

In dit onderzoek werd data van 96 startende studenten uit de opleiding Graduaat Programmeren onderzocht. De experimentele groep bestaat uit 63 deelnemers, de controlegroep uit 33.

Om na te gaan of de leeropbrengst in de experimentele groep minstens even groot is als in de controlegroep werd een T-toets uitgevoerd. De deelnemers uit de verschillende groepen zijn onafhankelijk van elkaar en de afhankelijke variabele leeropbrengst werd gemeten op intervalniveau. Een betrouwbaarheidsinterval (CI) van 95% op het steekproefgemiddelde werd gehanteerd, 95% van de steekproeven zal het populatiegemiddelde bevatten. Het significantieniveau werd vastgelegd op .05, wat een standaardwaarde is voor wetenschappelijk onderzoek en een evenwicht houdt tussen het risico op een type I – het onterecht aannemen van de hypothese - en type II fouten – het onterecht verwerpen van de hypothese. Een p-waarde kleiner .05 zal dus de hypothese dat er een significant

verschil is tussen de gemiddelden van de experimentele en controlegroep bevestigen, een p-waarde van .05 en hoger bevestigt de nulhypothese, wat zou aangeven dat er geen aantoonbaar verschil is tussen de gemiddelden. Bijlage E toont een screenshot uit het programma G*Power die aangeeft dat voor het gegeven aantal deelnemers een power van 0.95 wordt bereikt. De effectgrootte werd uitgedrukt met de waarde van Cohen's d. Bijlage F toont de kwalitatieve labels voor de effectgroottes voor de verschillende waarden van Cohen's d.

Voor het beantwoorden van de tweede deelvraag en het toetsen van de tweede hypothese, waarbij de gebruikservaringen van experimentele groep met de feedbacktool werden nagegaan en vergeleken via de benchmark, werden de gemiddeldes voor elke UEQ-schaal uit dit onderzoek vergeleken met de meegeleverde benchmarkgemiddeldes. De items werden gehercodeerd in één richting en de waarden 1 tot 7 werden omgerekend tot waarden -3 tot +3. De bevindingen werden gerapporteerd volgens de evaluatieschalen ingebed in de tool (excellent, goed, boven gemiddeld, onder gemiddeld of slecht). De interne consistentie van UEQ schalen werd berekend op de onderzoeksdata en vergeleken met de benchmarkdata.

Voor het toetsen van de derde onderzoekshypothese zijn de scores van de experimentele en controlegroep op de overeenkomstige, gehercodeerde items van de UEQ-survey met elkaar vergeleken door middel van T-toetsen. Ook hier werd het significatieniveau vastgesteld op .05 en werd de effectgrootte, Cohen's d, berekend.

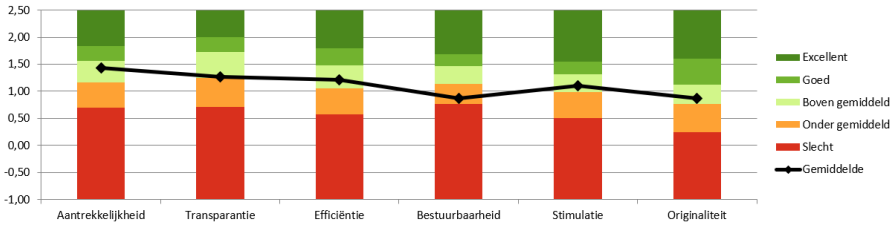
Resultaten

In dit onderzoek werd data van 96 studenten uit een opleiding Graduaat Programmeren aan en Vlaamse Hogeschool onderzocht. De experimentele groep bestaat uit 63 deelnemers, de controlegroep uit 33.

De belangrijkste bevindingen die uit het onderzoek naar voren komen zijn (1) startende studenten uit de opleiding Graduaat Programmeren in het Hoger Onderwijs behalen geen significant hogere of lagere leeropbrengst bij gebruik van een geautomatiseerde feedbacktool (EG) dan studenten die begeleid worden door een lector (CG), (2) de geautomatiseerde feedbacktool PluralSight wordt voor de UEQ-schalen aantrekkelijkheid, transparantie, efficiëntie, stimulatie en originaliteit boven gemiddeld en voor de schaal bestuurbaarheid onder gemiddeld gewaardeerd in vergelijking met de UEQ benchmarkresultaten, (3) de deelnemers uit de experimentele groep waarderen de studie ervaring op 11 van de 13 UEQ-items minstens even hoog als de deelnemers uit de controlegroep.

Gemiddeld behaalden de deelnemers uit de experimentele groep, gebruik makend van de geautomatiseerde feedbacktool PluralSight, een hogere leeropbrengst (M = .89, SE = .39) dan de deelnemers in de standaard onderwijssetting (M = .42, SE = .48) met begeleiding door één lector. Dit verschil, .47 CI [-.77, 1.70] was niet significant (t(71.87) = .75, p= .46) en de effectgrootte was triviaal (d = .16). Als antwoord op de eerste deelvraag kan dus gesteld worden dat de leeropbrengst niet significant hoger of lager is bij het gebruik van de geautomatiseerde feedbacktool dan bij begeleiding door een lector. De eerste hypothese, waarbij werd gesteld dat de leeropbrengst bij studenten die gebruik maken van de digitale begeleidingstool PluralSight minstens even groot is als bij de deelnemers die begeleiding kregen in een standaard onderwijssetting, kan worden bevestigd

Voor het beantwoorden van deelvraag 2 werden de scores voor de tool PluralSight voor de verschillende UEQ-schalen vergeleken met de benchmarkresultaten. De benchmarkdata werd door de UEQ-onderzoeksgroep ter beschikking gesteld (Schrepp et al., 2014) en maakt het mogelijk de gebruikte tool af te wegen ten opzichte van eerdere onderzoeksresultaten met andere tools. Bijlage F toont voor de UEQ schalen (aantrekkelijkheid, transparantie, efficiëntie, bestuurbaarheid, stimulatie en originaliteit) voor elk van de waardenschalen (slecht, onder gemiddeld, boven gemiddeld, goed en excellent) de bovengrens van de schaal, het gemeten gemiddelde en standaardfout van de onderzoeksdata. Figuur 2 visualiseert de waargenomen schaalgemiddeldes en de waardenschalen uit de benchmark. De waargenomen gemiddeldes op elke UEQ-schaal werden voor de tool PluralSight grafisch voorgesteld door de zwarte punten, verbonden door een lijn.



Figuur 2. Vergelijking van de UEQ-schaalgemiddelden voor PluralSight met de UEQ Benchmark van de experimentele groep. De score indeling van de benchmark wordt gevisualiseerd door de kleuren in balken. De zwarte lijn verbindt de schaalgemiddelden voor de tool PluralSight.

Op de schaal aantrekkelijkheid (n = 63, M = 1.44, SE = .09) scoort PluralSight boven gemiddeld – met ondergrens 1.17 en bovengrens 1.56 - in vergelijking met de benchmarkdata. Op de schaal transparantie (n = 63, M = 1.27, SE = .12) scoort PluralSight boven gemiddeld – met ondergrens 1.25 en bovengrens 1.73 - in vergelijking met de benchmarkdata. Op de schaal efficiëntie (n = 63, M = 1.21, SE = .12) scoort PluralSight boven gemiddeld – met ondergrens 1.05 en bovengrens 1.48 - in vergelijking met de benchmarkdata. Op de schaal bestuurbaarheid (n = 63, M = .87, SE = .12) scoort PluralSight onder gemiddeld – met ondergrens .77 en bovengrens 1.13 - in vergelijking met de benchmarkdata. Op de schaal stimulatie (n = 63, M = 1.10, SE = .09) scoort PluralSight boven gemiddeld – met ondergrens .99 en bovengrens 1.31 - in vergelijking met de benchmarkdata. Op de schaal originaliteit (n = 63, M = .87, SE = .07) scoort PluralSight boven gemiddeld – met ondergrens .77 en bovengrens 1.12 - in vergelijking met de benchmarkdata. De tweede hypothese kan dus bevestigd worden voor de schalen aantrekkelijkheid, transparantie, efficiëntie, stimulatie en originaliteit, maar wordt tegengesproken voor de schaal bestuurbaarheid.

Voor het nakijken van de interne consistentie van de schalen werd Cronbach's alpha berekend. De interne consistentie drukt uit in welke mate de items uit een schaal gerelateerd zijn als groep. We beschouwen een waarde groter dan .70 als aanvaardbaar, groter dan .80 als goed. Scoort de schaal onder .70 dan beschouwen we de interne consistentie als onaanvaardbaar. Tabel 2 toont de resultaten voor de UEQ-schalen. We zien dat de interne consistentie van de gebruikte schalen aantrekkelijkheid en transparantie goed is. Voor de schalen efficiëntie, bestuurbaarheid en stimulatie is de interne consistentie aanvaardbaar. Voor de schaal originaliteit is de interne consistentie onaanvaardbaar.

Tabel 2

UEQ-schaalgegevens van de experimentele groep, interne consistentie experimentele groep en benchmark.

	Experimentele groep (n = 63)				Benchmark (n = 18483)
Schaal	M	SD	95% CI	α	α
Aantrekkelijkheid	1.44	.73	[1.26 1.62]	.88	.81
Transparantie	1.27	.95	[1.04 1.51]	.82	.80
Efficiëntie	1.21	.96	[.98 1.45]	.76	.78
Bestuurbaarheid	.87	.98	[.63 1.11]	.70	.56
Stimulatie	1.10	.69	[.93 1.27]	.75	.79
Originaliteit	.87	.59	[.73 1.02]	.38	.70

Noot. M = gemiddelde; SD = standaarddeviatie; CI = confidence interval; α = Cronbach's alpha, maat voor de interne consistentie van een schaal.

Met opmerkingen [TS1]: 2 keer alfa in header. Wat is verschil?

Vergelijken we de berekende interne consistentie met de data uit de benchmark, dan is voor de schalen aantrekkelijkheid, transparantie, efficiëntie en stimulatie een score in de buurt van de benchmarkscore waar te nemen. Cronbach's alpha wijkt in de onderzoeksdata minder dan .10 af van de benchmarkdata. Voor de schaal bestuurbaarheid is de interne consistentie van de onderzoeksdata gevoelig hoger dan de benchmarkdata, voor de schaal originaliteit is deze gevoelig lager.

De derde deelvraag werd beantwoord door de 13 items van de UEQ-survey die door zowel de experimentele groep als de controlegroep werden beantwoord te vergelijken door middel van T-toetsen. Bijlage H toont de resultaten van deze T-toetsen voor elk van de onderzochte items van de UEQ survey. Uit deze resultaten blijkt dat voor 11 van de 13 onderzochte items met als positieve termen begrijpelijk, creatief, interessant, goed, eenvoudig, motiverend, volgens verwachting, efficiënt, overzichtelijk, praktisch en ordelijk, de experimentele groep gemiddeld een hogere score gaf dan de controlegroep. 2 Items met positieve termen plezierig en ondersteunend werden door de experimentele groep gemiddeld lager gescoord. Hoewel de bevindingen van slechts 4 items, met positieve termen eenvoudig, motiverend, overzichtelijk en ordelijk, statistisch significant blijken kunnen we uit deze resultaten toch afleiden dat de studenten die geautomatiseerde feedback van de tool PluralSight hebben ontvangen de studie-ervaring minstens even goed waarderen als bij begeleiding door een lector. De hogere scores voor de geautomatiseerde feedbacktool geven aan dat de tool PluralSight door de deelnemers gewaardeerd wordt. De opleiding kan concreet actie ondernemen om deze tool in te zetten.

Item nummer 1 uit de UEQ survey met negatieve term onplezierig en positieve term plezierig werd door de deelnemers uit de experimentele groep (n = 63, M = 1.38, SE = .11) lager gescoord dan door de deelnemers van de controlegroep (n = 33, M = 1.45, SE = .16). Dit verschil, -.07 CI [-.45, .30] was niet significant (t(94) = -.39, p = .70) en de effectgrootte was triviaal (d = -.08).

Item nummer 2 uit de UEQ survey met negatieve term onbegrijpelijk en positieve term begrijpelijk werd door de deelnemers uit de experimentele groep (n = 63, M = 1.44, SE = .16) hoger gescoord dan door de deelnemers van de controlegroep (n = 33, M = 1.09, SE = .27). Dit verschil, .35 CI [-.28, .99] was niet significant (t(55.20) = 1.11, p = .27) en de effectgrootte was zwak positief (d = .25).

Item nummer 3 uit de UEQ survey met negatieve term saai en positieve term creatief werd door de deelnemers uit de experimentele groep (n = 63, M = 1.14, SE = .14) hoger gescoord dan door de deelnemers van de controlegroep (n = 33, M = .76, SE = .26). Dit verschil, .39 CI [-.21, .98] was niet significant (t(51.62) = 1.29, p = .20) en de effectgrootte was zwak positief (d = .29).

Item nummer 7 uit de UEQ survey met negatieve term oninteressant en positieve term interessant werd door de deelnemers uit de experimentele groep (n = 63, M = 1.59, SE = .11) hoger gescoord dan door de deelnemers van de controlegroep (n = 33, M = 1.27, SE = .25). Dit verschil, .32 CI [-.23, .86] was niet significant (t(44.46) = 1.17, p = .25) en de effectgrootte was zwak positief (d = .27).

Item nummer 11 uit de UEQ survey met negatieve term belemmerend en positieve term ondersteunend werd door de deelnemers uit de experimentele groep (n = 63, M = 1.11, SE = .17) lager gescoord dan door de deelnemers van de controlegroep (n = 33, M = 1.52, SE = .16). Dit verschil, -.40 CI [-.87, .06] was niet significant (t(88.43) = -1.74, p = .86) en de effectgrootte was zwak negatief (d = -.35).

Item nummer 12 uit de UEQ survey met negatieve term slecht en positieve term goed werd door de deelnemers uit de experimentele groep (n = 63, M = 1.81, SE = .12) hoger gescoord dan door de deelnemers van de controlegroep (n = 33, M = 1.30, SE = .24). Dit verschil, .51 CI [-.02, 1.03] was niet significant (t(48.19) = 1.92, p = .06) en de effectgrootte was zwak positief (d = .44).

Item nummer 13 uit de UEQ survey met negatieve term complex en positieve term eenvoudig werd door de deelnemers uit de experimentele groep (n = 63, M = 1.11, SE = .14) hoger gescoord dan door de deelnemers van de controlegroep (n = 33, M = .24, SE = .36). Dit verschil, .87 CI [.08, 1.65] was significant (t(42.31) = 2.23, p = .03) en de effectgrootte was middelsterk positief (d = .52).

Item nummer 18 uit de UEQ survey met negatieve term demotiverend en positieve term motiverend werd door de deelnemers uit de experimentele groep (n = 63, M = 1.38, SE = .11) hoger gescoord dan door de deelnemers van de controlegroep (n = 33, M = .52, SE = .29). Dit verschil, .87 CI [.24, 1.49] was significant (t(41.81) = 2.79, p = .01) en de effectgrootte was middelsterk positief (d = .65).

Item nummer 19 uit de UEQ survey met negatieve term niet volgens verwachting en positieve term volgens verwachting werd door de deelnemers uit de experimentele groep (n = 63, M = 1.03, SE = .15) hoger gescoord dan door de deelnemers van de controlegroep (n = 33, M = .48, SE = .29). Dit

verschil, .55 CI [-.10, 1.20] was niet significant (t(50.92) = 1.69, p = .10) en de effectgrootte was zwak positief (d = .38).

Item nummer 20 uit de UEQ survey met negatieve term efficiënt en positieve term inefficiënt werd door de deelnemers uit de experimentele groep (n = 63, M = 1.02, SE = .15) hoger gescoord dan door de deelnemers van de controlegroep (n = 33, M = .67, SE = .23). Dit verschil, .35 CI [-.17, .87] was niet significant (t(94) = 1.33, p = .19) en de effectgrootte was zwak positief (d = .28).

Item nummer 21 uit de UEQ survey met negatieve term verwarrend en positieve term overzichtelijk werd door de deelnemers uit de experimentele groep (n = 63, M = 1.17, SE = .17) hoger gescoord dan door de deelnemers van de controlegroep (n = 33, M = .18, SE = .32). Dit verschil, .99 CI [.27, 1.72] was significant (t(50.11) = 2.74, p = .01) en de effectgrootte was middelsterk positief (d = .62).

Item nummer 22 uit de UEQ survey met negatieve term praktisch en positieve term onpraktisch werd door de deelnemers uit de experimentele groep (n = 63, M = 1.21, SE = .13) hoger gescoord dan door de deelnemers van de controlegroep (n = 33, M = .82, SE = .26). Dit verschil, .39 CI [-.19, .97] was niet significant (t(47.63) = 1.35, p = .19) en de effectgrootte was zwak positief (d = .31).

Item nummer 23 uit de UEQ survey met negatieve term rommelig en positieve term ordelijk werd door de deelnemers uit de experimentele groep (n = 63, M = 1.65, SE = .14) hoger gescoord dan door de deelnemers van de controlegroep (n = 33, M = .52, SE = .31). Dit verschil, 1.14 CI [.44, 1.83] was significant (t(44.67) = 3.31, p = < .01) en de effectgrootte was middelsterk positief (d = .76).

Conclusie en discussie

Dit onderzoek bouwt verder op eerder onderzoek naar gebruik van automatische feedbacktools. Hier werd soms een positief verband gesteld (Kenny & Pahl, 2009; McCartney, 2019; Stevenson & Phakiti, 2014) tussen het gebruik van deze tools en de studieprestatie, waarbij in ander onderzoek (Wieling & Hofman, 2010) geen verband kon worden vastgesteld. Onderzoek naar de relatie tussen het gebruik van een geautomatiseerde feedbacktool op de leeropbrengst en de studie-ervaring kan deze inzichten verder vormgeven. De beperking van de deelnemers tot slechts één opleiding en de beperkte deelnemersgroep noopt echter tot voorzichtigheid bij veralgemenen van deze resultaten. De vraag rijst of de resultaten uit deze studie, waarbij de focusgroep studenten in een IT-opleiding zijn, kan geëxtrapoleerd worden naar andere opleidingen. De deelnemers uit deze studie zijn niet enkel vertrouwd met het gebruik van de computer, het is veelal hun interesse of zelf passie. Het is goed mogelijk dat andere groepen minder open staan voor het gebruik van deze geautomatiseerde tools. Verder onderzoek op dit vlak is dus wenselijk.

Als antwoord op deelvraag1 concluderen we dat deelnemers die de geautomatiseerde feedback tool gebruikten geen significant hogere of lagere leeropbrengst behaalden dan deelnemers uit de controlegroep. De verwachting werd uitgesproken dat de leeropbrengst bij het gebruik van de geautomatiseerde feedbacktool minstens even groot zou zijn als in een traditionele onderwijssetting. Deze hypothese kan door dit onderzoek worden bevestigd. Deze bevestiging halen we uit de redenering dat dit onderzoek de hypothese niet tegenspreekt, er is geen reden om aan te nemen dat het gebruik van de geautomatiseerde feedbacktool de leeropbrengst zou verminderen. Dat de onmiddellijke, elaboratieve feedback van de geautomatiseerde feedbacktool een positieve invloed op het leerproces kan hebben, zoals eerder aangetoond (Chase & Houmanfar, 2009; J. Hattie, 2008; Kenny & Pahl, 2009), wordt door dit onderzoek niet tegengesproken noch bevestigd, verder onderzoek is dus wenselijk.

Dit resultaat dient gezien te worden in een context waarbij de geboden feedback in de experimentele en controlegroep niet dezelfde is. In de experimentele groep werd door de tool geautomatiseerd directe, elaboratieve feedback voorzien. De controlegroep, die feedback ontving door een lector, ontving de feedback niet steeds direct, de studenten moeten soms wachten tot de lector beschikbaar is. Bijkomend onderzoek naar de invloed van het verhogen van de kwantiteit en timing van de geboden feedback in de controlegroep is hierbij wenselijk. Inzetten van meerdere instructors of het verkleinen van het aantal deelnemers kan deze setting mogelijk maken.

Dat PluralSight een wereldwijd platform is, gebruikt door ruim 700.000 developers, ontwikkeld door een professioneel team, leidde alvast tot positieve verwachtingen omtrent de gebruikerservaring. De UEQ-survey voor het meten van de gebruikerservaring bevestigt de gestelde hypothese dat de tool PluralSight even goed of beter gewaardeerd wordt dan andere tools in de benchmarkresultaten voor 5 UEQ-schalen: aantrekkelijkheid, transparantie, efficiëntie, stimulatie en originaliteit. Voor 1 schaal, bestuurbaarheid, zien we de gestelde hypothese niet bevestigd. Wanneer we dieper ingaan op de items van de schaal bestuurbaarheid, met positieve termen voorspelbaar, ondersteunend, vertrouwd en ‘volgens verwachting’ kan opgemerkt worden dat deze items wellicht beter zullen scoren naarmate de studenten de tool meer aanwenden voor het maken van oefeningen. Voor de studenten waren de interactieve oefeningen met PluralSight een eerste ervaring. Bij vervolgonderzoek met dezelfde studenten lijkt het aannemelijk dat de items voorspelbaar, vertrouwd, volgens verwachting uit de schaal bestuurbaarheid hoger zullen scoren. Ook andere schalen bevatten items zoals begrijpelijk, snel, eenvoudig, aangenaam, efficiënt, overzichtelijk, waarvan kan verwacht worden dat die bij herhaald gebruik hoger zullen scoren. Natuurlijk is het voor bevestiging hiervan wenselijk dit vervolgonderzoek uit te voeren. Het lijkt dus echter aannemelijk te stellen dat de score op de gebruikerservaring van de digitale feedbacktool PluralSight bij herhaald gebruik zal stijgen. Dit vervolgonderzoek kan ook aantonen of items lager gaan scoren bij herhaald gebruik. Kandidaten hiervoor zijn de items met

positieve termen spannend, origineel, aantrekkelijk. Het lijkt me aannemelijk te stellen dat naarmate de tool vaker wordt ingezet deze minder spannend, origineel en aantrekkelijk wordt. Vanzelfsprekend moet hier opgemerkt worden dat deze bevinding ook gelden voor de andere tools uit de benchmark, maar er is omtrent het herhaald gebruik en de invloed op de gebruikerservaring hiervan geen inzicht in de benchmarkdata (Schrepp et al., 2017). Gezien het inzetten van de tool PluralSight binnen de opleiding mogelijk is voor andere lessen Javascript, maar ook voor andere programmeertalen, zoals C# en SQL, zullen de studenten deze tool veelvuldig kunnen gebruiken. Het bevestigen van deze tweede hypothese motiveert alvast om de tool verder in de opleiding Graduaat Programmeren in te zetten.

De derde hypothese, die stelt dat de geautomatiseerde feedback even goed of beter gewaardeerd wordt als feedback door een lector in een grote studentengroep, wordt bevestigd voor 11 van de 13 geselecteerde items uit de UEQ-survey. De positieve termen van de items die in de experimentele groep beter scoren zijn begrijpelijk, creatief, interessant, goed, eenvoudig, motiverend, volgens verwachting, efficiënt, overzichtelijk, praktisch en ordelijk. 2 Items met positieve termen plezierig en ondersteunend werden door de experimentele groep gemiddeld lager gescoord. Hoewel de bevindingen van slechts 4 items, met positieve termen eenvoudig, motiverend, overzichtelijk en ordelijk, statistisch significant blijken kunnen we uit deze resultaten toch aannemen dat de studenten die geautomatiseerde feedback van de tool PluralSight hebben ontvangen de studie-ervaring minstens even goed waarderen als bij begeleiding door een lector. De derde hypothese wordt dus niet voor alle onderzocht items bevestigd maar wordt door dit onderzoek ook niet tegengesproken. Deze bevindingen bouwen verder op de eerdere vaststelling dat digitale annotaties goed bruikbaar zijn voor studenten (Ryan et al., 2019), de hogere score van de experimentele groep op de items begrijpelijk en overzichtelijk bevestigen de vastgestelde verzuchting van studenten dat feedback van lectoren soms moeilijk te begrijpen is (Ferguson, 2011; Hounsell et al., 2008; Hyland, 2013). Het streven naar het geven van begrijpelijke en overzichtelijke feedback is dus op zich voor het opleidingsteam ook een meteen een werkpunt. Op het gebied van de studie-ervaring kent het verder inzetten van de digitale feedbacktool PluralSight in de opleiding ondersteuning door deze studie.

Een globaal aandachtspunt voor inzicht in deze resultaten is de inhoud van het studiemateriaal voor de deelnemers. De proefpersonen volgden een instructie Javascript, waarbij enkel de basisconcepten werden behandeld. Dergelijke leerinhoud leent zich wellicht beter in het gebruik van geautomatiseerde tools dan inhoud die zich focust op een groter redeneervermogen. De mogelijke fouten die de deelnemers konden maken, en dus de nood aan feedback, speelden zich vooral af op het vlak van syntax en semantiek, minder op het vlak van het inzichtelijke of het conceptuele van de leerstof. Het gebruik van deze digitale feedbacktool zou dus ook positief kunnen zijn bij het inoefenen van niet complexe vaardigheden binnen andere vakgebieden, bijvoorbeeld het inoefenen van

basisconcepten wiskunde, natuurkunde of woordenschat (van den Broek et al., 2019). Zo kan begeleidingstijd vrijgemaakt worden bij de docent, die dan besteed kan worden aan het verwerven van complexe vaardigheden. Dus ook op het vlak van inzetbaarheid binnen andere domeinen dan programmeren is verder onderzoek zeker aangewezen.

Inzichten in de relatie tussen het gebruik van geautomatiseerde tools en de leeropbrengst en studie-ervaring kunnen onderwijsbeslissingen helpen sturen. Het opleidingsteam kan met deze nieuwe inzichten aan de slag om het leerproces voor de studenten zo efficiënt mogelijk te maken, binnen een maatschappij waarin de digitale evolutie nauwelijks bij te houden is. Deze studie kan de beslissingen wetenschappelijke grond geven en vermijden dat hervorming en digitalisering enkel als doel worden gezien en niet als middel.

Referenties

Abran, A., Khelifi, A., Suryn, W., & Seffah, A. (2003). Usability Meanings and Interpretations in ISO Standards. *Software Quality Journal*, *11*(4), 325–338. https://doi.org/10.1023/A:1025869312943

Agricola, B. T., Prins, F. J., & Sluijsmans, D. M. A. (2020). Impact of feedback request forms and verbal feedback on higher education students' feedback perception, self-efficacy, and motivation. *Assessment in Education: Principles, Policy & Practice*, *27*(1), 6–25. https://doi.org/10.1080/0969594X.2019.1688764

Altherwi, M., & Gravell, A. (2019). A Large-Scale Dataset of Popular Open Source Projects. *Journal of Computers*, *14*(4), 7.

Birenbaum, M., Breuer, K., Cascallar, E., Dochy, F., Dori, Y., Ridgway, J., Wiesemes, R., Editor, & Nickmans, G. (2006). A learning Integrated Assessment System. *Educational Research Reviewuk (J. Ridgway)*, *1*, 61–67. https://doi.org/10.1016/j.edurev.2006.01.001

Chase, J. A., & Houmanfar, R. (2009). The Differential Effects of Elaborate Feedback and Basic Feedback on Student Performance in a Modified, Personalized System of Instruction Course. *Journal of Behavioral Education*, *18*(3), 245–265. https://doi.org/10.1007/s10864-009-9089-2

Delialioğlu, Ö. (2012). Student Engagement in Blended Learning Environments with Lecture-Based and Problem-Based Instructional Approaches. *Journal of Educational Technology & Society*, *15*(3), 310–322. afh.

Digital Devils. (2020). *Random Code Generator*. https://www.randomcodegenerator.com/nl/generate-codes

Driscoll, M. P. (2014). *Psychology of Learning for Instruction: Pearson New International Edition*. Pearson Education Limited.

Ehrenberg, R. G., Brewer, D. J., Gamoran, A., & Willms, J. D. (2017). Class Size and Student Achievement: *Psychological Science in the Public Interest*. http://journals.sagepub.com/doi/10.1111/1529-1006.003

Exeter, D. J., Ameratunga, S., Ratima, M., Morton, S., Dickson, M., Hsu, D., & Jackson, R. (2010). Student engagement in very large classes: The teachers' perspective. *Studies in Higher Education*, *35*(7), 761–775. https://doi.org/10.1080/03075070903545058

Ferguson, P. (2011). Student perceptions of quality feedback in teacher education. *Assessment & Evaluation in Higher Education*, *36*(1), 51–62. https://doi.org/10.1080/02602930903197883

Graham, C. R. (2006). *Handbook of blended learning.* Pfeiffer.

Hamel, R., Côté, K., Matte, A., Lepage, J.-F., & Bernier, P.-M. (2019). Rewards interact with repetition-dependent learning to enhance long-term retention of motor memories. *Annals of the New York Academy of Sciences*, *1452*(1), 34–51. https://doi.org/10.1111/nyas.14171

Hattie, J. (2008). *Visible Learning: A Synthesis of Over 800 Meta-Analyses Relating to Achievement*. Taylor & Francis.

Hattie, John, & Gan, M. (2011). Instruction based on feedback. *Handbook of Research on Learning and Instruction*, 249–271.

Hattie, John, & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Heusser, A. C., Awipi, T., & Davachi, L. (2013). The ups and downs of repetition: Modulation of the perirhinal cortex by conceptual repetition predicts priming and long-term memory. *Neuropsychologia*, *51*(12), 2333–2343. https://doi.org/10.1016/j.neuropsychologia.2013.04.018

Hounsell, D., McCune, V., Hounsell, J., & Litjens, J. (2008). The quality of guidance and feedback to students. *Higher Education Research & Development*, *27*(1), 55–67. https://doi.org/10.1080/07294360701658765

Hyland, K. (2013). Student perceptions of hidden messages in teacher written feedback. *Studies in Educational Evaluation*, *39*(3), 180–187. https://doi.org/10.1016/j.stueduc.2013.06.003

ISO. (2016). *ISO/IEC 25023:2016*. ISO. https://www.iso.org/cms/render/live/en/sites/isoorg/contents/data/standard/03/57/35747.html

Jordan, P. W., Thomas, B., McClelland, I. L., & Weerdmeester, B. (1996). *Usability Evaluation In Industry*. CRC Press.

Kenny, C., & Pahl, C. (2009). Intelligent and adaptive tutoring for active learning and training environments. *Interactive Learning Environments*, *17*(2), 181–195. https://doi.org/10.1080/10494820802090277

Kinshuk, P., & Russell, D. (2000). A multi-institutional evaluation of intelligent tutoring tools in numeric disciplines. *Educational Technology & Society*, *3(4): 66–74*.

Kulhavy, R. W. (1977). Feedback in Written Instruction. *Review of Educational Research*, *47*(2), 211–232. https://doi.org/10.3102/00346543047002211

Laugwitz, B., Held, T., & Schrepp, M. (2008). *Construction and Evaluation of a User Experience Questionnaire* (Vol. 5298). https://doi.org/10.1007/978-3-540-89350-9_6

Lee, I. (2019). Teacher written corrective feedback: Less is more. *Language Teaching*, 1–13. https://doi.org/10.1017/S0261444819000247

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human–Computer Interaction*, *7*(1), 57–78. https://doi.org/10.1080/10447319509526110

Luscombe, C., & Montgomery, J. (2016). Exploring medical student learning in the large group teaching environment: Examining current practice to inform curricular development. *BMC Medical Education*, *16*, 184–184. PubMed. https://doi.org/10.1186/s12909-016-0698-x

McCartney, M. (2019). Debunking an active-learning myth. *Science*, *363*(6425), 361. https://doi.org/10.1126/science.363.6425.361-e

Mulryan-Kyne, C. (2010). Teaching large classes at college and university level: Challenges and opportunities. *Teaching in Higher Education*, *15*(2), 175–185. https://doi.org/10.1080/13562511003620001

Osgood, C. E., Suci, G., & Tannenbaum, P. H. (1957). *The Measurement of Meaning*. University of Illinois Press.

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education*, *35*(3), 277–289. https://doi.org/10.1080/02602930903541007

Ryan, T., Henderson, M., & Phillips, M. (2019). Feedback modes matter: Comparing student perceptions of digital and non-digital feedback modes in higher education. *British Journal of Educational Technology*, *50*(3), 1507–1523. https://doi.org/10.1111/bjet.12749

Sadler, D. R. (2010). Beyond feedback: Developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, *35*(5), 535–550. https://doi.org/10.1080/02602930903541015

Schrepp, M. (2015). *User Experience Questionnaire Handbook*. https://doi.org/10.13140/RG.2.1.2815.0245

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2017). Construction of a Benchmark for the User Experience Questionnaire (UEQ). *International Journal of Interactive Multimedia and Artificial Intelligence*. https://www.ijimai.org/journal/sites/default/files/files/2016/12/ijimai20174_4_5_pdf_94297.pdf

Schrepp, M., Hinderks, A., & Thomaschewski, J. (2014). *Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios*. 383–392. https://doi.org/10.1007/978-3-319-07668-3_37

Shute, V. J. (2008). Focus on Formative Feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Singh, R., Gulwani, S., & Solar-Lezama, A. (2012). *Automated Feedback Generation forIntroductory Programming Assignments*.

Stevenson, M., & Phakiti, A. (2014). The effects of computer-generated feedback on the quality of writing. *Feedback in Writing: Issues and Challenges*, *19*, 51–65. https://doi.org/10.1016/j.asw.2013.11.007

Stobart, G. (2008). *Testing Times: The Uses and Abuses of Assessment*. Routledge. https://doi.org/10.4324/9780203930502

Sweller, J. (2019). Cognitive load theory and educational technology. *Educational Technology Research and Development*. https://doi.org/10.1007/s11423-019-09701-3

Theisen, K. J. (2019). Programming languages in chemistry: A review of HTML5/JavaScript. *Journal of Cheminformatics*, *11*(1), 11. https://doi.org/10.1186/s13321-019-0331-1

Valcke, M. (2017). *Onderwijskunde als ontwerpwetenschap*. Academia Press.

van den Broek, G. S. E., Segers, E., van Rijn, H., Takashima, A., & Verhoeven, L. (2019). Effects of elaborate feedback during practice tests: Costs and benefits of retrieval prompts. *Journal of Experimental Psychology: Applied*, *25*(4), 588–601. https://doi.org/10.1037/xap0000212

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Review of Educational Research*, *85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Van Merriënboer, J. J. G., & Kirschner, P. A. (2013). *Ten steps to complex learning (2nd ed.): A systematic approach to four-component instructional design.* Routledge.

Wieling, M. B., & Hofman, W. H. A. (2010). The impact of online video lecture recordings and automated feedback on student performance. *Computers & Education*, *54*(4), 992–998. https://doi.org/10.1016/j.compedu.2009.10.002

Winstone, N., & Carless, D. (2019). *Designing Effective Feedback Processes in Higher Education: A Learning-Focused Approach*. Routledge. https://doi.org/10.4324/9781351115940

Bijlage A: de vragen van de pretest

De pretest test de voorkennis van de deelnemers, tabel A1 geeft je een zicht op de gestelde vragen. Er wordt gepolst naar correcte syntax en semantiek voor het gebruik van de programmeertaal Javascript binnen de leerstofonderdelen variabelen, functies en selectie.

Tabel A1

De vragen van de pretest. De antwoordmogelijkheden, het correcte antwoord en het gewicht van de vraag zijn voorzien.

Vraag	Mogelijkheden	Correct Antwoord	Gewicht

Welke uitvoer genereert volgende code?
`var isOk = true;`
`isOk = !isOk;`
`if(!isOk) console.log("a");`
`else console.log("b");`
	• a
	• b
	• NaN
	• SyntaxError
		a	2

Welke uitvoer genereert volgende code?
`let countStudents = 10;`
`double(countStudents);`
`console.log(countStudents)`
 	
`function double(countStudents)`
`{`
`  return countStudents * 2;`
`}`
	• 12
	• 10
	• 20
	• NaN
		10	3

Welke uitvoer genereert volgende code?
`let count = 10;`
`console.log(testSimpleCondition(count));`
 	
`function testSimpleCondition(input) {`
`  if(input >= 10)`
`  return input * 2;`
`  return input * 5;`
`}`
	• NaN
	• 10
	• 20
	• 50
		20	2

Welke uitvoer genereert volgende code?	• 3	17	3

```
var number = 16;
console.log(testComplexCondition(number));

function testComplexCondition(input) {
   if(input < 15 && input >3)
      return input++;
   return ++input;
}
```

- 15
- 16
- 17

Welke uitvoer genereert volgende code?

```
let name = 'John Doe';
try {
if(verifyCondition(name))
   console.log("Verified - ok");
else
   console.log("Verified - NOT ok");
} catch(error) {
   console.log("NOT verified");
}

function verifyCondition(text) {
   if(text.substring(0,1) == 'X'
|| (text.length > 7 && text.substring(0,1) == "J"))
      return true;
   return false;
}
```

- Verified – ok
- Verified – NOT ok
- NOT verified
- SyntaxError

Verified - ok 5

Bijlage B: de vragen van de posttest

De posttest test de kennis van de leerinhoud na de treatmentles, tabel B1 geeft je een zicht op de gestelde vragen.

Er wordt gepolst naar correcte syntax en semantiek voor het gebruik van de programmeertaal Javascript binnen het nieuwe leerstofonderdeel over de iteratie: while en for loops.

Tabel B1

De vragen van de posttest. De antwoordmogelijkheden, het correcte antwoord en het gewicht van de vraag zijn voorzien.

Vraag	Mogelijkheden	correct	Gewicht
Een while loop kent in het signatuur het gebruik van het sleutelwoord while, gevolgd door, tussen ronde haakjes: • Een verplicht argument dat leidt tot een boolean value • Een optioneel argument dat leidt tot een boolean value • Een verplicht argument dat leidt tot een integer value • Een optioneel argument dat leidt tot een integer value	Zie vraag	1	1
Een for-lus kent drie argumenten, je kan deze als volgt omschrijven 1. **for**(uit te voeren na de loop**;** initialisatie**;** voorwaarde controleren) 2. **for**(voorwaarde controleren**;** initialisatie**;** uit te voeren na de loop) 3. **for**(uit te voeren na de loop**;** voorwaarde controleren**;** initialisatie) 4. **for**(initialisatie**;** voorwaarde controleren**;** uit te voeren na de loop)	Zie vraag	4	2

Welke uitvoer genereert volgende code?
`let studentsCounter = 1;`
`let studentsInClass = 36;`
`let studentsArrived = 0;`
 	
`while(studentsCounter < studentsInClass) {`
`  studentsArrived++;`
`}`
	• Geen uitvoer
	• 1
	• 35
	• 36
		Geen uitvoer	2

```
console.log(studentsArrived);
```

Welke uitvoer genereert volgende code?
`let waarde = 10;`
 	
`while(waarde > 5) {`
`  waarde -= 2;`
`}`
 	
`console.log(waarde);`
	• 10
	• 6
	• 4
	• SyntaxError
		4	2

Welke uitvoer genereert volgende code?
`let som = 10;`
`for(let teller = 50; teller < 56; teller++)`
`{`
`  som += 5;`
`}`
`console.log(som);`
	• NaN
	• 10
	• 35
	• 40
		40	3

`let aantalFietsen = 4;`
`let aantalAutos = 3;`
`let bedrag = 0;`
 	
`let huidigeFiets = 1;`
`while(huidigeFiets < aantalFietsen) {`
`  huidigeFiets++;`
`  bedrag += 20;`
`}`
`for(let huidigeAuto = 1;`
`huidigeAuto < aantalAutos;`
`huidigeAuto++)`
`{`
`  bedrag += 50;`
`}`
`console.log(bedrag);`
	• 20
	• 50
	• 160
	• 180
			5

Bijlage C: UEQ surveyitems voor de experimentele groep

#	Begrip	Tegengestelde	UEQ-schaal
1	Plezierig	Onplezierig	Aantrekkelijkheid
2	Onbegrijpelijk	Begrijpelijk	Transparantie
3	Creatief	Saai	Originaliteit
4	Makkelijk te leren	Moeilijk te leren	Transparantie
5	Waardevol	Inferieur	Stimulatie
6	Vervelend	Spannend	Stimulatie
7	Oninteressant	Interessant	Stimulatie
8	Onvoorspelbaar	Voorspelbaar	Bestuurbaarheid
9	Snel	Langzaam	Efficiëntie
10	Origineel	Conventioneel	Originaliteit
11	Belemmerend	ondersteunend	Bestuurbaarheid
12	Goed	Slecht	Aantrekkelijkheid
13	Complex	Eenvoudig	Transparantie
14	Afstotend	Aantrekkelijk	Aantrekkelijkheid
15	Gebruikelijk	Nieuw	Originaliteit
16	Onaangenaam	Aangenaam	Aantrekkelijkheid
17	Vertrouwd	Niet vertrouwd	Bestuurbaarheid
18	Motiverend	Demotiverend	Stimulatie
19	Volgens verwachting	Niet volgens verwachting	Bestuurbaarheid
20	Inefficiënt	Efficiënt	Efficiëntie
21	Overzichtelijk	Verwarrend	Transparantie
22	Onpraktisch	Praktisch	Efficiëntie
23	Ordelijk	Rommelig	Efficiëntie
24	Aantrekkelijk	Onaantrekkelijk	Aantrekkelijkheid
25	Aardig	Onaardig	Aantrekkelijkheid
26	Conservatief	Innovatief	Originaliteit

Noot. # = UEQ-itemnummer.

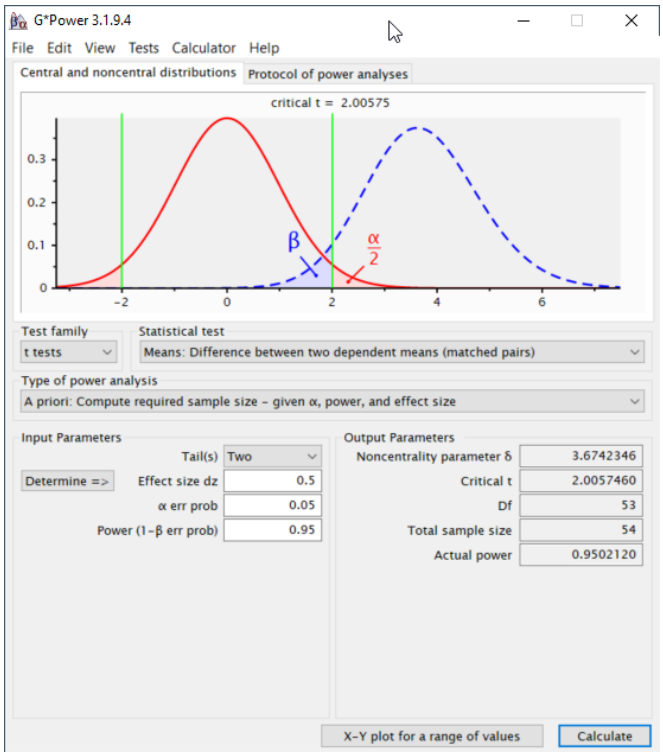
Bijlage D: UEQ surveyitems voor de controlegroep

#	Grensbegrip	Tegengestelde
1	Plezierig	Onplezierig
2	onbegrijpelijk	begrijpelijk
3	Creatief	Saai
7	Oninteressanet	Interessant
11	Belemmerend	ondersteunend
12	Goed	Slecht
13	Complex	Eenvoudig
18	Motiverend	Demotiverend
19	Volgens verwachting	Niet volgens verwachting
20	Inefficiënt	Efficiënt
21	Overzichtelijk	Verwarrend
22	Onpraktisch	Praktisch
23	Ordelijk	Rommelig

Noot. # = UEQ-itemnummer.

Bijlage E: T-test – G*Power

Figuur E1 toont een screenshot uit de applicatie G*Power met de poweranalyse voor de T-test. Het aantal gewenste deelnemers voor experimentele groep en controlegroep werd berekend.



Figuur E1. G*Power analyse voor de T-test.

Bijlage F: Cohen's d - kwalitatieve labels voor de samenhang

Cohen's d	Samenhang
Kleiner dan -1.3	Zeer sterk negatief
Tussen -1.3 en -0.8	Sterk negatief
Tussen -0.8 en -0.5	Middelsterk negatief
Tussen -0.5 en -0.2	Zwak negatief
Tussen -0.2 en 0.2	Triviaal
Tussen 0.2 en 0.5	Zwak positief
Tussen 0.5 en 0.8	Middelsterk positief
Tussen 0.8 en 1.3	Sterk positief
Groter dan 1.3	Zeer sterk positief

Bijlage G: UEQ benchmark en vergelijk met de experimentele groep

UEQ-schaal	Bovengrenzen voor de waardenschalen uit benchmark					Experimentele groep	
	Slecht	Onder gemiddeld	Boven gemiddeld	Goed	Excellent	M	SE
Aantrekkelijkheid	.7	1.17	1.56	1.83	3	1.44	.09
Transparantie	.71	1.25	1.73	2	3	1.27	.12
Efficiëntie	.57	1.05	1.48	1,8	3	1.21	.12
Bestuurbaarheid	.77	1.13	1.46	1.69	3	.87	.12
Stimulatie	.5	.99	1.31	1.55	3	1.10	.09
Originaliteit	.25	.77	1.12	1.61	3	.87	.07

Noot. M = gemiddelde; SE = standaardfout

Bijlage H: resultaten T-toetsen voor vergelijk UEQ-items

	Experimentele groep (n = 63)		Controlegroep (n = 33)		T-Test					
nr	M	SE	M	SE	MD	95% CI	df	t	p	d
1	1.38	.11	1.45	.16	-.07	[-.45, .30]	94	-.39	.70	-.08
2	1.44	.16	1.09	.27	.35	[-.28, .99]	55.20	1.11	.27	.25
3	1.14	.14	.76	.26	.39	[-.21, .98]	51.62	1.29	.20	.29
7	1.59	.11	1.27	.25	.32	[-.23, .86]	44.46	1.17	.25	.27
11	1.11	.17	1.52	.16	-.40	[-.87, .06]	88.43	-1.74	.86	-.35
12	1.81	.12	1.30	.24	.51	[-.02, 1.03]	48.19	1.92	.06	.44
13	1.11	.14	.24	.36	.87	[.08, 1.65]	42.31	2.23	.03	.52
18	1.38	.11	.52	.29	.87	[.24, 1.49]	41.81	2.79	.01	.65
19	1.03	.15	.48	.29	.55	[-.10, 1.20]	50.92	1.69	.10	.38
20	1.02	.15	.67	.23	.35	[-.17, .87]	94	1.33	.19	.28
21	1.17	.17	.18	.32	.99	[.27, 1.72]	50.11	2.74	.01	.62
22	1.21	.13	.82	.26	.39	[-.19, .97]	47.63	1.35	.19	.31
23	1.65	.14	.52	.31	1.14	[.44, 1.83]	44.67	3.31	< .01	.76

Noot. nr = nummer van het item in de UEQ-survey; M = gemiddelde; SE = standaardfout; MD= verschil tussen de gemiddeldes; CI = confidence interval; df = vrijheidsgraden; p = probabiliteit voor statistische significantie ; d = Cohen's d